CA Journal of Cultural Analytics

Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction

Matthew Wilkens

11.01.16

Peer-Reviewed By: Scott Selisker

Clusters: Genre

Journal ISSN: 2371-4549 DOIs: 10.22148/16.009 (CA) and 10.7910/DVN/EXPXYT(Dataverse)

Is "literary fiction" a useful genre label in the post-World War II United States? In some sense, the answer is obviously yes; there are sections marked "literary fiction" on Amazon, in bookstores, and on Goodreads, all of which contain many postwar and contemporary titles. Much of what is taught in contemporary fiction classes also falls under the heading of literary fiction, even if that label isn't always used explicitly. On the other hand, literary fiction, if it hangs together at all, may be defined as much by its (or its consumers') resistance to genre as by its positive textual content. That is, where conventional genres like the detective story or the erotic romance are recognizable by the presence of certain character types, plot events, and narrative styles, it is difficult to find any broadly agreeable set of such features by which literary fiction might be consistently identified.

Literary fiction, then, might be understood as an anti-genre genre. Perhaps. But my own suspicion is that literary fiction functions in genre terms calibrated to the tastes of an upper-middlebrow literary market that values the relative invisibility of genre and is resistant to reading in generic terms. Indeed, as Mark McGurl and Caren Irr have separately shown in recent books, it is possible to discern a handful of subclasses within the broad category of "serious" postwar literature and to find links between genre and non-genre writing, though the broad coherence of literary fiction itself remains in some doubt.

If my hunch about the high-level contours of literary fiction in the twentieth century is correct, it should be possible to identify one or more groups of contemporary literary fiction that hang together at least as closely as do the texts that are more conventionally called genre fiction. While the specific features marking out these literary genres will be unique - just as war novels differ in form and content from vampire tales - their functional codification will not. Identifying such a genre or genres would allow literary scholars to understand the dynamics of the contemporary novel in part through keywords similar to those used to treat mass-market genres (including "community," "formula," and "identification"). A generic understanding of literary fiction would also provide further motivation for the ongoing revision of the literary canon by pointing to the implicit standardization of much literary fiction in contrast to the formal and topical diversity that scholars generally seek.

Genre is, however, a difficult concept to pin down, one that has been treated unevenly and for a range of ends in literary criticism, sociology, and the cultural marketplace. For critics in the mold of Northrop Frye, genre is a formal category having to do with the "radical of presentation" through which plot-level events are portrayed; for Gérard Genette, it involves primarily content-level differences within formally defined modes; for Raymond Williams, genre is part of a "social language" that unites distinct aspects of the social and material processes that make up a cultural situation.[1] For Amazon and Netflix, genres cohere around a mixture of thematic content, style of presentation, and communities of

[1]For Frye's formalist version of genre, see especially essay four of his *Anatomy of Criticism* (Princeton: Princeton UP, 1957). For Genette's more content-oriented approach, see *The Architext: An Introduction*. Translated by Jane E. Lewin, forward by Robert Scholes, (Berkeley: U California P, 1992). Williams's socioaesthetic formulation is in part three, chapter six of *Marxism and Literature* (Oxford: Oxford UP, 1977).

consumption, which explains why "religious" and "gay and lesbian" are two of the largest Amazon fiction genres and why both "international films" and "witty TV comedies featuring a strong female lead" are Netflix genres.

Within literary fiction proper, McGurl's recent work on the influence of creative writing programs has helped to distinguish the related subgenres of "technomodernism," "high-cultural pluralism," and "middle-class minimalism," each of which shades not only into the other two, but also into the larger, longer-standing genres of science fiction, detective fiction, and life writing.[2] McGurl's generic anatomy is valuable, and we will see evidence of the genealogies he identifies in the results below, especially those linking tendencies within conventionally identified postwar literary fiction to extraliterary and pulp genres. Irr's work on the varieties of what she calls the "geopolitical" novel presents a detailed analysis of internationalism in twenty-first century U.S. fiction along lines that might be understood as similarly generic, if necessarily more specialized than any theorization genre systems as such.[3] Both McGurl and Irr highlight the inseparability of social and textual factors in their treatments of genre, a fact that serves to highlight the increasing salience of cultural information to genre construction as publishing and readership alike have opened to more diverse participation over the course of the twentieth century.

Even if critics agreed on one of these definitions, however, there would remain the task of identifying and categorizing the features through which it is manifested in specific texts. That is, how exactly does a reader become aware that she's reading a detective story or a neoliberal allegory? To what can she point in the text or in her situation of consumption that signals generic membership?

The research presented here is designed to answer three linked questions about genre and literary fiction in the twentieth-century U.S. First, what is a reasonable set of features through which to distinguish texts belonging to different genres? Second, are there major genres or genre-like groupings that have not been previously identified within the most widely collected and circulated novels of the period? That is, can we build a model of genre in twentieth-century fiction without knowing in advance all the genres we expect to find and without having read most of the books in question, and if so, does such a model suggest novel literary affinities? Third, to what extent does the subset of literary fiction that is the subject of much academic and intellectual interest cohere in generic terms? If it does cohere strongly, what can we make of the observed pattern of inclusion and exclusion?

To answer these questions requires three things in turn: a (very) rough theory of genre, a corpus of relevant texts, and one or more strategies for translating that theory of genre into specific judgments about group membership for the corpus texts. The working theory is rough indeed: I call a genre a set of texts that resemble one another in subject matter, style of presentation, setting, and, to a modest extent, circumstances of production such as publication date and author identity. Details follow below, but the idea is that genre is constituted by a range of family resemblances, most (but not all) of which are textual and all of which are susceptible to identification, if only by proxy, and to imperfect quantification. What this definition loses in theoretical sophistication, it gains in its range of potential application; it is a framework intended to cover many colloquial uses of the term "genre" in relation to the novel and to be extensible in future work to other mass-cultural aesthetic forms including film, television, music, and the like.

A full account of the corpus and the methods used to analyze it follow immediately below, but first, a brief summary of the findings. Computational analyses using unsupervised machine learning techniques on more than 8,500 American novels published during the long twentieth century suggest that the working theory of genre offered here can be made functional and interpretable in a range of cases. The results indicate an increasing fluidity of genre in the post-World War II period, when it becomes more difficult to tease apart many novels outside certain highly codified forms (especially detective fiction and war stories). In the absence of extra-textual information, we find an intriguing continuity between early-century regionalist writing and postwar science fiction. The addition of even a very modest amount of historical and social data weakens this association, but suggests that canonical and near-canonical contemporary fiction by male writers including Kurt Vonnegut, Saul Bellow, Walker Percy, Richard Yates, Don DeLillo, Tim O'Brien, Stephen King, Philip K. Dick, John Updike, and James Michener both shares a high degree of text-level similarity and is strongly atypical of postwar fiction overall, a result that argues for the continuing value of revisions to the literary-academic canon. Gender also appears to be a good predictor of at least some additional generic divides, especially within what we would ordinarily call commercial genre fiction after 1945.

[2] Mark McGurl, *The Program Era: Postwar Fiction and the Rise of Creative Writing* (Cambridge: Harvard UP, 2009).
[3] Caren Irr, *Toward the Geopolitical Novel: U.S. Fiction in the Twenty-First Century* (New York: Columbia UP, 2013).

Corpus and methods

The corpus to be analyzed is sampled from a list of the most frequently held novels and novel-like texts by American authors published between 1880 and 1990 as cataloged by WorldCat. The research set comprises 8,580 volumes totaling nearly one billion words, distributed bimodally with peak holdings around 1900 and the 1980s. Note that the corpus does not contain novels published after 1990 and is therefore not well suited to addressing questions of genre in very contemporary U.S. fiction.

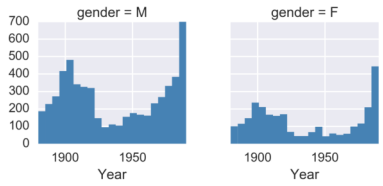

Figure 1. Distribution of volumes in the research corpus by date of publication and author gender.

Bibliographic information about the texts, including author genders and ethnicities and dates of original publication, was collected by hand. Slightly less than a third of the texts are by women; only a tiny subset (about 3%) of those for which the author's ethnicity is known are by non-white writers. The range of sources of these imbalances - which include everything from direct discrimination in the publishing industry and in library acquisitions, to inequalities of economic and educational opportunity for authors, to the changing demographics of readership - are currently under active investigation. Nevertheless, the corpus is representative of twentieth-century American fiction as codified by the current and past practices of many academic and cultural institutions. It is obviously not identical to the set of all U.S. novels published in the twentieth century, but it is much less canonical than the selection principle might lead one to expect. It contains, after all, nearly 9,000 novels. Most of them are obscure.

Some additional observations on the composition of the corpus may help readers to better understand its coverage. The single most-held book is Toni Morrison's *Beloved*. Somewhat counterintuitively, books by minority writers seem to be held either very widely or not at all; the small percentage of included volumes by non-white authors is disproportionately concentrated in the top 20% of books ranked by library holdings.[4] Also prominent are late works by canonical authors (Faulkner's *The Reivers* and Hemingway's *Islands in the Stream*, for instance); historical fiction in the vein of Jean M. Auel - who wrote *Clan of the Cave Bear* and a series of similar novels - and James A. Michener; and middlebrow titles by the likes of Anne Tyler, John Irving, and Pat Conroy. But most of the books in the corpus are unknown to me and to the literary scholars with whom I have consulted.[5] Note that, because library holding data are not used in the computational model, none of these works is weighted in any special way beyond the fact of its inclusion in the corpus.

As explained at the outset, we are treating genre as a matter of similarity between texts, rather than as a question of conformity to any existing set of paradigmatic examples. The open-ended nature of this task is what dictates an unsupervised computational learning method. Briefly, unsupervised classification methods differ from supervised methods in that they do not assume any "correct" classification output, hence they do not rely on training data to select and weight feature inputs. The latter approaches - *supervised* (as opposed to *unsupervised*) methods - can be attractive to the extent that their performance is directly measurable over a set of gold-standard classifications, typically produced, in literary cases, by experts in the field. When the performance of a supervised method is low, users can tune parameters and inputs and observe any resulting improvement in classification accuracy. The right-wrong nature of supervised output is also an important check on what we might call interpretive overfitting, that is, the human ability to tell a just-so story about the results of almost any classification task.[6] But users do need to be able to supply training data that looks, for the most part, like the corpus as a whole and includes the range of desired, known categories. The task, in that case, is to find the right features and techniques to reproduce the judgments embedded in the training data.

[4] My thanks to Richard So for this observation.

[5] A corpus listing and related project data are available in the *CA* dataverse.

[6] Supervised learning tasks can, of course, also include confidence measures among their outputs, including the possibility of assignment to more than one category. The difference between supervised and unsupervised techniques is thus emphatically not a matter of objective vs. interpretable results (which would be to confuse the nature of interpretation). For an example of using supervised classification accuracy as an input to critical analysis, see Ted Underwood, "The Life Cycles of Genres," *CA: Journal of Cultural Analytics* (23 May 2016).

In the present instance, however, the premise is specifically that we do not know what the range of correct classifications in the full corpus might be. Here, then, we define our task as identifying aggregate volume-level similarity across a range of features that capture significant aspects of genre as defined above and as used in many (though, of course, not all) critical treatments of the term. The features in question fall under a small set of headings:

1. Subject matter, measured in the present case by topic-modeled word frequencies.
2. Style, form, and diction, measured by volume-level statistics including reading-level score, verb fraction, text length, etc.
3. Setting and location, assessed via geolocation extraction and geosimilarity measures.
4. A limited range of extra-textual features, including publication date and author gender. Extra-textual features are excluded from some of the models below in order to compare text-only results to more expansive conceptions of generic affinity.

This is a small set, but many major components of actually existing critical discourse about genre can be identified under one of its broad headings. That said, the goal is to select a set of potentially useful features to use as algorithmic inputs; evaluation will depend on the extent to which those features help to reproduce textual groupings that resemble critical judgments about generic membership. If the output is good, it won't matter a great deal, for present purposes, that one might prefer a slightly different set. If the output is bad, it won't matter how abstractly well-justified any of the features might be. What follows, then, is less a detailed argument for the specific areas and features deployed than it is a summary of the choices involved.

Subject matter is straightforwardly the most important of the set. Detective fiction is about detectives and murders and the police. Romance, in the modern sense, is about desire and bodies and love. Historical novels involve objects and situations adapted (if not notably accurately) to a specific time and place. This is the main way readers know that they have encountered one of these genres, and thematic considerations retain, as we have seen, a certain pride of place in generic systems from the theoretical to the commercial.

To quantify content and subject matter, I have used latent Dirichlet allocation topic modeling.[7] The approach is to build a 200-topic model of the corpus via MALLET, which model is then reduced to 20 principal components, each consisting of a different relative weight attached to all 200 modeled topics.[8] This dimensionality reduction ensures that over-split and highly correlated topics are treated together as content features and produces, for the most part, relatively interpretable components for the model. Twenty of the 29 total modeled features used in the full model - 69% - are thus devoted to specific content rather than to form, geography, or bibliographic data.

Despite the prominence of subject matter, stylistic and formal features also matter. This is true in part due to the final inseparability of form from content; thematic questions are always also formal questions. We know this already on theoretical grounds and it is easy to see directly in the case of the noir, for example, where the short, declarative sentence is a near requisite to generic membership. The issue may be even more pressing with respect to identifiably "literary" fiction, which is picked out, on the anti-generic reading, less by its content than by the register of its diction and style of its narration. In fact, the results below suggest precisely that certain types of post-1945 literary fiction are distinguished by a combination of modern content with typically prewar stylistic features. The genre of the encyclopedic novel, moreover, is perhaps more about length than Edward Mendelson or I would care to admit.[9] Erotica and romance are distinguished only in part by the explicitness of their content; the way in which their stories are told matters, too.

The stylistic features deployed here are four: text length as measured by word count; grade-level reading score, which combines average sentence length with average syllable count per word; the fraction of the word count made up of terms that entered the English language before 1150, as described by Underwood and Sellers, which provides a proxy measure for the "elevation" of the text's diction (higher fractions of early terms correspond to lower, less Latinate diction); and the fraction of all words in the text that are verbs, a measure of narrative balance between action and description.[10]

Setting and location are relevant to genre in two ways. First, there is the longstanding consideration of public and domestic

[7] M. David Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993-1022 ; Andrew McCallum Kachites, "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu. 2002.

[8] The 20 principal components (PCs) capture 52% of the underlying variance in the 200 topics.

[9] Mendelson's insightful and influential treatment of the encyclopedic novel is "Encyclopedic Narrative: From Dante to Pynchon," *MLN* 91 (1976): 1267-1275. My own analysis of literary encyclopedism is found in chapter four of *Revolution: The Event in Postwar Fiction* (Baltimore: Johns Hopkins UP, 2016).

[10] Ted Underwood and Jordan Sellers, "The Emergence of Literary Diction," *Journal of Digital Humanities* 1.2 (2012).

space associated with differently gendered narratives. And a love story set in the wilderness but deploying the conventions of the manor is a comedy as much as a romance. Second, there is the matter of national and geographic divisions that don't necessarily track either author identity or textual content alone, as between, say, the Holmesian and hard-boiled detective story, or between historical fiction of the Second World War and the distinct genre of counter-factual history (the latter containing a mashup of modern and mid-century geographies). The transposition of an established generic form to a different geographic context may also represent part of the difference between, say, the realisms of Thomas Hardy, Charles Dickens, and Sarah Orne Jewett. Spatial issues are perhaps not the primary determinants of genre, but they have a bearing that is not always well captured by thematic and formal features.

Three input features are thus devoted to geography. These are assembled from extracted geolocation data produced via named entity recognition and automated geocoding (with limited hand correction) as described in Wilkens.[11] The data - which ranges from continents and oceans to nations, cities, parks, and individual buildings - is then aggregated by nation and by U.S. state.[12] The distribution of the resulting aggregates across the corpus texts are consolidated into a TF-IDF matrix, which is then reduced from 268 dimensions (corresponding to the 268 unique nations and U.S. states in the data) to 3 principal components, which are used as features in the models. I note in passing that these components can be succinctly, if quite loosely, summarized as "New York," "California and not Europe," and "Europe and not the American South or Midwest."

Extra-textual features are perhaps the most fraught inclusion in any evaluation of genre. If we understand genre labels as attached to groups of texts that resemble one another, we might insist that such resemblance be measured according to textual features alone. But to do so would seem to ignore a large part of both critical and readerly practice. We speak of women's writing and African-American literature in ways that, while not identical to genre, are not altogether distinct, and we would often insist that knowledge concerning authorial identity is at least as relevant as the more obviously generic fictional or nonfictional status of a text. The publication date of a book also matters so long as we retain historicist aspirations for our scholarship. As Borges taught us, Pierre Menard wrote a much different work than did Cervantes; *Werther*, written today, is related to the memoir boom in a way that Goethe's original is not. These two extra-textual features - date of publication and author gender - are used as inputs to the full model, but are excluded in certain cases indicated below to allow a comparison between that model and one trained on textual features alone.

There are thus 29 (or 27, in the text-only case) specific features included in the analyses below, with the number of feature types in each of the four classes corresponding to the relative weight of that class in the model. The features and class weights are summarized in table 1.

Class	Details	#	Weight (full)	Weight (text-only)
Content	Principal components of 200-topic model	20	0,69	0,74
Form	Length, reading level, diction, verb fraction	4	0,14	0,15
Geography	Principal components of geographic TF-IDF	3	0,10	0,11
Extra-textual*	Publication date, author gender	2	0,07	–

Table 1. Summary of features used for unsupervised learning. (* indicates features excluded from the text-only model.)

There is room for productive disagreement concerning both the identity and the details of these features, the determination of which represents a series of interpretive and practical choices guided by the preceding elaboration of genre. Additional textual and extra-textual data could certainly be valuable. Commercial information including sales price, copies sold, and the market orientation of the issuing press or imprint - all adjusted for changes over more than a century of cultural history - might help to track generic filiations, as could details about cover imagery and reader reception networks, though the incorporation of large amounts of such data could serve to reproduce too strongly the existing genre categories that we seek to interrogate and would certainly pose serious collection difficulties. Alternate approaches to assessing content and formal similarity might be developed. Greater representation of racial and ethnic diversity in the underlying corpus would allow important additional questions to be addressed. Nevertheless, the attempt has been in every case to construct features that, while feasible to collect and tractable to compute, are maximally tied to the social

[11] Matthew Wilkens, "The Geographic Imagination of Civil War-Era American Fiction," *American Literary History* 25.4 (2013): 803-840.

[12] Recall that the corpus consists entirely of U.S. fiction. More than 60% of named location occurrences fall within the United States. Indeed, New York and locations therein are the single most frequently occurring area.

and formal issues presented by the problem of genre. But support for their appropriateness will come ultimately in the form of the classification outputs built on top of them.

Some of these features are correlated, though the dimensionality reduction performed on the content- and geographic-type data minimizes correlations within those sets. Figure 2 shows feature correlations (Pearson product-moment coefficients) in the full data set. Notable correlations (positive and negative) exist between publication date and the first few topic components; between diction score and the third topic component; and between diction score and reading level. None of these correlations is unexpected. Further analysis of the interactions between features follows in the discussion of the models, below.

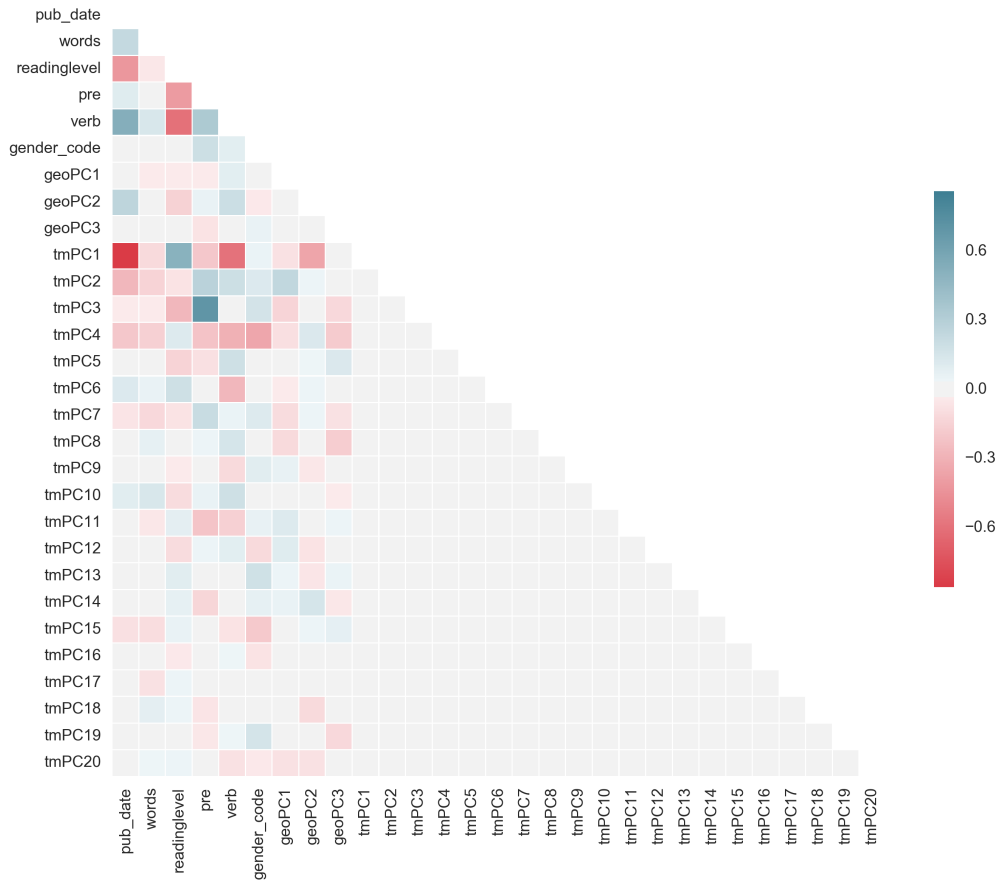

Figure 2. Correlations among input variables.

The features are normalized to mean zero and variance 1, then reduced to ten dimensions via PCA to produce suitable density given the corpus size. Finally, I perform two types of clustering over the feature matrix. The first is classic *k*-means, which takes as an input the desired number of output clusters and returns an assignment to one of those clusters for each of the input texts. The assignment is determined by finding the set of cluster centers that minimizes the average distance between those centers and the data points closest to each of them. *K*-means is an older statistical technique, pioneered in the 1950s and '60s, but it is easy to perform and to interpret, and is a suitable choice when the number of clusters is known in advance (or can be derived from heuristic analysis of the results).

The second clustering technique, DBSCAN, is more recent and works on a different principle. It identifies regions of high relative density among points in the feature space, assigning the points that fall within any such region to a cluster.[13]

[13] Ester, Martin et al, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Nois," In *Proceedings of 2nd International*

DBSCAN is thus robust to noise (outlying points are assigned to no cluster at all) and, within reason, to low-density feature spaces (in which it will detect only the few high-density regions), a fact that will be important in the analysis of generic patterns in post-1945 fiction.

Feature and model validation

As always with unsupervised learning, the question "is it working?" has to be answered with interpretive interventions. To test the validity of the procedure, I generated a 50-cluster model including extra-textual data over the full corpus via *k*-means and examined the resulting assignments for a test set of 302 novels by 25 authors.[14] The number of included books per author in the test set varied from as few as two (Jack London) to as many as 37 (William Dean Howells); the median was 10 volumes per author. These books and writers are ones for which reasonably canonical genre assignments seemed possible, as indicated in figure 3, but even here, there is room for meaningful disagreement about the details. Authors were selected for inclusion in the test set in part on the basis of the relative generic consistency of their texts so as to minimize instances in which texts by a single author are expected to assort to different clusters.

What we are looking for in the algorithmic assignments isn't necessarily the precise arrangement depicted in figure 3 (which, after all, conflates texts and authors in a way that implies a false sense of perfect generic consistency within any author's oeuvre), but a plausible alignment of similar authors and texts with one another both within and across clusters. That is to say, most authors should have most of their works assigned to a small number of clusters, and most clusters should contain works by generically similar authors.

Conference on Knowledge Discovery and Data Mining (KDD-96), (AAAI Press: 1996), 226-231.

[14]The number of clusters was set well above the "best" result (k = 15-20) as measured by silhouette score to minimize agglomeration errors, albeit at the expense of splitting errors. The idea is that it is relatively easy to identify and merge erroneously split clusters, but very difficult to disaggregate a small number of heterogeneous clusters.

Figure 3. Expected generic groups in a test set of 302 novels by 25 authors.

The output across 50 clusters is difficult to present concisely, because it is necessarily more diverse than the limit case of figure 3. Seventeen of the 25 authors (68%) have at least half of their works assigned to (different) single clusters from among the 50 clusters available, a good result given the decision to split clusters aggressively by selecting a large *k*. The authors whose works are more diversely categorized are, for the most part, "literary" writers not usually identified as having worked in a single genre, including John Updike (on whom, more below), Joyce Carol Oates, John Steinbeck, William Dean Howells, and Willa Cather. This result lends some support to the anti-generic view of literary fiction, though the density-based results below complicate such a conclusion. Surprisingly, Philip K. Dick and Patricia Highsmith - both more solidly genre authors - likewise see their work spread across several clusters, a result illuminated by the text-only DBSCAN findings below.

Also relevant is the range of authors included within each cluster and, more specifically, the number of clusters that produce surprising (hence potentially incorrect or newly informative) groupings. Of the 14 clusters that contain multiple texts by each of two or more test-set authors (that is, clusters that are more mixed in their authorial membership), most are as expected: Howells with Edith Wharton and Henry James; Isaac Asimov with Robert A. Heinlein and Dick; Zane Grey with Louis L'Amour; Don DeLillo with Philip Roth, Richard Yates, and Updike. But two are unexpected, including the genre writers Mary Higgins Clark and Highsmith grouped with Oates and Yates; and Elmore Leonard and Dick with John Steinbeck.

It's hard to say, in the absence of a full critical treatment of the books in question, whether these are outright errors or previously unremarked lines of filiation between some of the twentieth century's most widely read authors. In the first

case, the Clark-Highsmith-Oates-Yates group, the model appears to be picking up features in the selected texts related to comparatively lowbrow postwar fiction set in eastern U.S. locations: middling reading-level scores, lack of elevated diction, a relatively high frequency of verbs, a general preference for contemporary topical content, and a preponderance of eastern U.S. settings at the expense of both Europe and the western U.S. This combination of features is not unique in the test set, but a handful of texts by these four writers match one another quite closely. The Leonard-Dick-Steinbeck cluster is less certain, though they share some topical content and, more prominently, both a preference for western settings and fairly simple sentence structures.

These two surprising clusters from among the fifty generated do little to undermine confidence in the validity of the model. They may in fact increase it, since some degree of lack of fit between computational models and existing categories is surely to be expected and can be explained in concrete terms. The same process performed on text-only data produces broadly similar outputs. On the whole, the *k*-means results suggest that the selected feature sets are capable of producing generic clusters that broadly resemble existing critical judgments, an important step toward validating the approach and the choices that enable it.

Unnatural kinds

There's no such thing as a natural kind, of course, and the methodological details above should make clear that there are multiple, active critical interventions involved in both the setup and analysis of unsupervised learning techniques. But the *k*-means model isn't ideally suited to identifying an unknown number of rare, high-density clusters embedded within large collections.[15] Yet such clusters may be exactly the ones that literary critics seeking to understand genre in the twentieth century would most like to locate, since they are the potential cores of a modified genre system.

As noted in the methods section above, the DBSCAN algorithm is a good fit for this problem. We can use density-based clustering to select only those regions of the feature space that contain groups of highly similar texts and to exclude large swaths of the corpus that are not so closely aligned.[16] In principle, this should leave us with a set of clusters that resemble genres according to our flexible definition.

We begin by performing DBSCAN clustering on the full corpus using the set of text-only features (that is, excluding publication data and author gender). This isn't the best match for our most robust articulation of genre, which clearly does involve social and historical factors. But by beginning with text-only data, we establish a baseline against which to evaluate the more complete data set - including the specific influence of introducing extra-textual information into the model - while simultaneously probing for strictly textual similarities that might be interesting in their own right.

The results of DBSCAN clustering on text-only features are shown in figure 4. The method assigns 657 total volumes (just under 8% of the corpus) to 8 distinct clusters.[17]

[15]The *k*-means results, in fact, are notable for the relative uniformity of the size of the output clusters, a fact related to the data's predominantly even distribution across high-dimensional feature space.

[16]To use a geographic metaphor, you might think of *k*-means as answering the question "to what province does this town belong"; every town will be assigned to one province or another. DBSCAN answers the question "is this town part of a significant urban agglomeration?"

[17]The large number of outlier texts in the DBSCAN output indicates a sparse population of texts in the feature space. Recall that all clustering operations were performed on a dimension-reduced version of the input feature set (27 dimensions, in the present case, reduced to 10 via PCA). Sparsity is desirable here because it allows us to identify only those subsets of the corpus that resemble one another very closely.

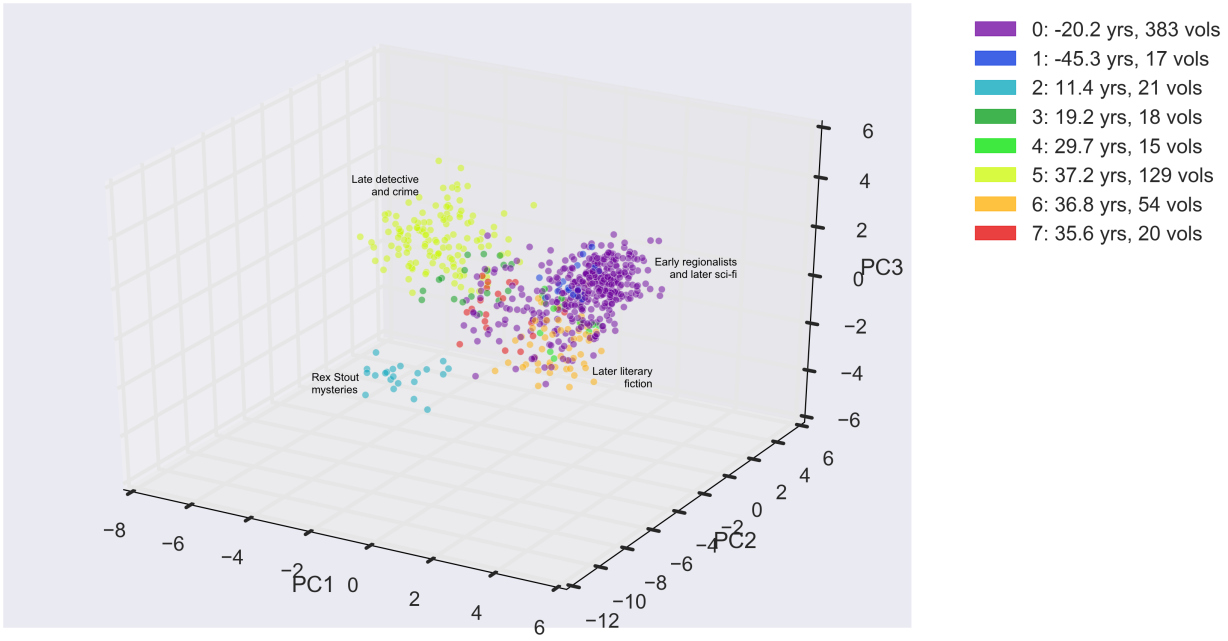

Figure 4. Eight "core" clusters identified by density-based analysis using the text-only feature set, arranged from earliest average publication date (purples and blues) to latest (yellows and reds) and projected into three-dimensional component space. Cluster dates are calculated relative to the average for all listed clusters (1943.0), which is later than the full corpus mean publication date (1936.6).

Three features are notable in the output. First, as Ted Underwood has also found in recent computational work on literary genre, we observe that detective fiction shows strong internal consistency, with the result that it is comparatively easy to distinguish from other genres.[18] This fact is most clearly visible in clusters 2, 3, and 5. The former comprises (almost) exclusively mysteries by the prolific mid-century writer and activist Rex Stout, the serial and esoteric nature of whose work likely explains its relative independence.[19] Cluster 3 contains hard-boiled detective and crime fiction by Mickey Spillane, George Harmon Coxe, Donald E. Westlake, and others. The larger cluster 5 includes texts by many of the leading practitioners of the genre, including Elmore Leonard, Sue Grafton, Marcia Muller, Charles Willeford, and Stuart M. Kaminsky, as well as a handful of detective-adjacent authors such as Philip K. Dick and Robert Ludlum. It's a notably recent cluster (average publication date, 1980) and a very predominantly male-authored one, Grafton and Muller's inclusion notwithstanding. In content, the texts use very little of the older, more descriptive vocabulary collected in the first principal component of the topic model, leaning more heavily on components 2, 4, and (especially) 5, each of which skews later and toward action, interior settings, and the material trappings of detective work (guns, phones, cars, detectives, police). Component 5 also includes some sentimental terms ("love," "heart," "God," and "death" give a flavor), in line with the suggestion that detective fiction, for all its association with calculation and masculine emotional reserve, also functions through an important if covert use of feeling.[20]

The second notable feature in figure 4 is the large cluster 0, which skews toward novels published in the late nineteenth and early twentieth centuries. It contains many important regionalist writers of the period, including Kate Chopin, Hamlin Garland, Booth Tarkington, Rebecca Harding Davis, George Washington Cable, and Mark Twain, as well as texts by less strongly regionalist-identified authors such as Sinclair Lewis, Edith Wharton, William Dean Howells (represented by a single volume, among his 34 included in the corpus), Katherine Anne Porter and an early volume of James A. Michener's. It is also, however, home to an unexpectedly large amount of postwar science fiction: three volumes by Ursula Le Guin, two from Philip K. Dick, five by L. Ron Hubbard, many more by less well-known writers, plus thrillers by Michael Crichton, Patricia Highsmith, and Robert Ludlum.

Cluster 0 is distinguished by form and setting more strongly than are other clusters. Its books tend toward fairly average

[18] Ted Underwood, "The Life Cycle of Genres".

[19] Cluster 2 in figure 4 contains a single non-Stout volume: *The Bilbao Looking Glass*, also a mystery, by Charlotte MacLeod.

[20] For a recent, genealogical treatment of emotion in detective and crime fiction, see Leonard Cassuto's *Hard-Boiled Sentimentality: The Secret History of American Crime Stories* (Columbia UP, 2009).

reading level scores, which, given the marked inverse correlation between reading level and publication date, means that the earlier volumes have somewhat low reading level values relative to others published during their time, while the later entries are comparatively difficult in the postwar era. These books also generally have lower verb fractions and less pre-1150 diction: they are (comparatively) heavy on description rather than action and use more Latinate vocabulary than their contemporaries. Most strikingly, they make little use of any of the three geographic components, meaning that they avoid the locations that dominate most fiction in the corpus. This makes sense as a further covert link between regionalist and science fiction, both of which genres are typically and importantly concerned with obscure or imaginary places outside those treated in other types of literature.

Even the small amount of more literary fiction that finds its way into cluster 0 fits the mold: John Okada's *No-No Boy* and Gil Scott-Heron's *The Nigger Factory* combine regionally distinctive elements with the slightly obscure diction and parabolic structure characteristic of science fiction. Indeed, a potential link between regionalist, minoritarian, and science fiction across the twentieth century is likely to be an area of rewarding investigation, one that extends work by Fredric Jameson, Mark McGurl, and the handful of critics who have addressed science fiction as a specifically racialized form.[21]

Finally, cluster 6 stands out as the model's most obviously literary postwar agglomeration. It contains exclusively volumes published from the fifties through the eighties and is marked, as expected for contemporary texts, by little use of topic component 1 (older, sentimental vocabulary). The books lean more heavily on components 6 and 10 (domestic life in urban settings), as well as on geographic component 1 (New York and the eastern U.S.), while avoiding geo component 3 (Europe). They tend to be longer than average, especially for postwar texts, but, as is common for more recent fiction, not notably complex at the sentence level. It is in this cluster that we find Saul Bellow, Stanley Elkin, Wallace Stegner (writing in his non-Western mode), Paul Theroux, Leon Uris, Jane Smiley, and, tellingly, Marge Piercy's classic, New York-based social dystopia *Woman on the Edge of Time*. This is another instance in which we see speculative fiction assorting with mainstream literary fiction, though here without the broad historical span that characterized cluster 0. On the whole, cluster 6 appears to identify the strain of familial, domestic, memoir-like fiction, often with a satirical edge, that has risen to particular prominence in the postwar period.

Still, we should note that many of the corpus texts that critics would typically identify as canonically literary - from Henry James and Willa Cather to Thomas Pynchon and Joyce Carol Oates - are absent from cluster 6 and from the text-only model as a whole. The DBSCAN algorithm, in combination with the selected set of text-only features, produces a model that identifies classical genre fiction - especially the detective story, mysteries, crime fiction, and certain types of science fiction - quite strongly, along with prewar regionalist writing that includes many of the most prominent American authors of the late nineteenth and early twentieth centuries. It suggests affinities between earlier regionalism and later science fiction and minoritarian writing, though the limits of the corpus curtail our ability to explore race and regionalism more fully.

Adding social data

As useful as is the text-only model, it specifically excludes basic social and historical information that we have good reason to believe rightly influences existing treatments of genre. Some of this extra-textual information makes its way into the results above by the textual back door, since there are correlations between certain text-level features and the excluded data (as shown in figure 3). Nevertheless, it is valuable to recalculate and evaluate the model over the full data set, both to reflect more closely the multifaceted theory of genre offered at the outset and to assess the impact on the model as a whole of adding small amounts of non-textual information.

Figure 5 summarizes the output of the DBSCAN algorithm over the full data set, including publication dates and author genders (where known). The algorithm assigns 685 total volumes (8% of the corpus) to 11 non-overlapping clusters.

[21] On science fiction and race, see especially André Carrington's recent André M. *Speculative Blackness: The Future of Race in Science Fiction* (Minneapolis: U of Minnesota P, 2016) and DeWitt Kilgore's *Astrofuturism: Science, Race, and Visions of Utopia in Space* (Philadelphia: U of Pennsylvania P, 2003).

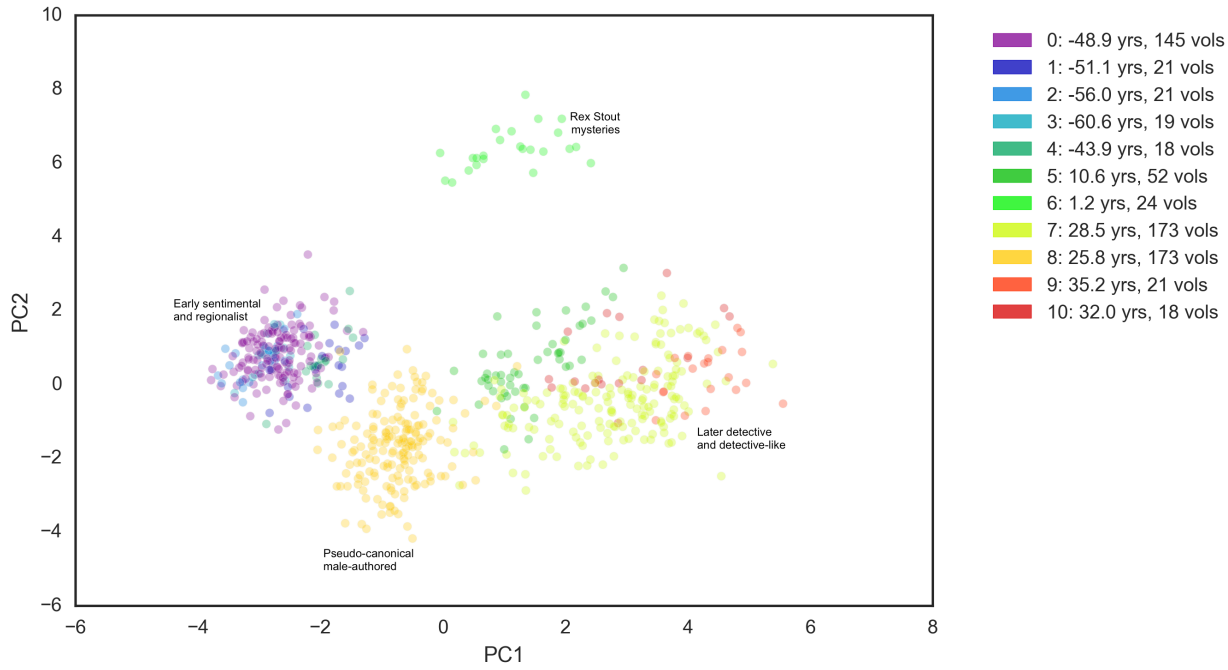

Figure 5. Eleven clusters identified by density-based analysis using the full feature set, arranged from earliest average publication date (purples and blues) to latest (yellows and reds) and projected into two-dimensional component space. Cluster dates are calculated relative to the average for all listed clusters (1951.8), which is later than the full corpus mean publication date (1936.6) and than the mean publication date of volumes in the text-only output (1943.0, figure 4). Two dimensions are used here, rather than the three in figure 4, to maximize readability in each case.

Among these 11 clusters, a few trends stand out. For one thing, the clusters mirror the bimodal publication date distribution of the corpus itself; there is a set of clusters centered in the late nineteenth and early twentieth century, and another set in the 1970s and '80s. There's almost nothing in the middle (other than Rex Stout), a result that mirrors the text-only output.[22] And recall that, although publication date does enter directly into this version of the model, it's only very lightly weighted (about 3%) as a feature in its own right. Books at the extreme ends of the corpus's historical range seem to supply some of the most distinctive formations.

Also notable is the amount of detective and crime fiction, again corresponding closely to the text-only result. Fully four of the clusters - together containing 115 volumes - fall almost exclusively within that genre. Two of the four (numbers 5 and 6 in figure 5) trend toward the early postwar period, centered around 1960 and containing texts by Mickey Spillane, Earl Stanley Garner, and - again in a cluster all his own - Rex Stout. The other two (numbers 9 and 10) are relatively small, come later, and are split by both geography and gender. Cluster 10 is entirely male-authored and New York-oriented, with heavy weights on component 1 of the geographic feature (recall that the geographic components describe, in loose overview, New York, California, and Europe). Cluster 10 is also, however, high on simple terms, generally of low reading score, and with topical content that runs toward the sentimental and romantic. Its authors are largely unknown, as is perhaps not surprising for a group producing short, easy-reading, sentimental genre fiction. The other contemporary detective cluster isn't strictly a detective cluster at all; it's built around the mysteries of Sue Grafton's alphabet series and the work of Marcia Muller. These are books set in the American West and that, while heavy on murder and mystery content, lack the quantity of direct reference to police procedures that helps to identify detective fiction proper. Together, clusters 9 and 10 are most directly comparable to cluster 5 in the text-only model; adding extra-textual information appears to reveal a potential split in postwar detective fiction not only between male and female-authored texts, but also between the subgenres of detection and mystery.

The large contemporary cluster, number 7, also bears mentioning. Its 173 volumes are exclusively male-authored (or by

[22]Note that the average per-cluster publication dates supplied in the legends of figures 4 and 5 are calculated independently relative to the mean dates of the volumes in each output set; because these means are different (the average publication date of volumes in figure 5 is 8.8 years later than figure 4), the cluster ages in figures 4 and 5 are not directly comparable.

authors of unknown gender) and share a significant investment in martial themes, both in war narratives and in accounts of crime. The writers are mostly obscure, though there are a handful of titles by Dick, Elmore Leonard, and a single entry from Robert Ludlum. Recalling that author gender accounts for just over 3% of the total feature weights, it is striking that these clusters separate so cleanly along gender lines, though we recall having observed a similar, if less stark, division in the text-only output.

Taken together, these detective, mystery, and crime clusters help to indicate the kind of objects and relations on which the method continues to perform well, just as it did in the absence of extra-textual information: tightly typified novels written close together in time, sharing a distinctive geography, and often segregated by author gender to a significant extent. It seems fair, again, to call this genre fiction in the most conventional sense. These density-identified groups collectively are also meaningful outliers, containing as they do fewer than 10% of the full corpus - a fact that will become especially relevant below.[23]

Our intuition that we've landed on a genre-finding method is strengthened by several of the earliest groupings in the full-data model. Cluster 1, for instance, appears to map directly onto turn-of-the-last-century sentimental fiction by women: short, relatively easy reading that is low on action (as measured by verb fraction) and loads heavily on component 1 of the topic model (containing love, faces, eyes, and absolutely nothing of police or offices). Interestingly, though, among the mostly little-known female authors in the cluster we find Kate Chopin, Paul Lawrence Dunbar, Hamlin Garland, and Edith Wharton's *Ethan Frome*. It is not a shock to see these books linked to the sentimental tradition, though Wharton's book is something of an exception in the existing criticism and may profit from revisitation in light of this link.

The larger early cluster, number 0, is more difficult to characterize. It is heavily male, contains sentimental content at levels only slightly below cluster 1, and similarly privileges long, descriptive, Latinate sentences. It's notably regionalist, with many books loading negatively on all three of the geographic components, suggesting a focus on locations outside the mainstream of American fiction. This tendency is reinforced by the group's few canonical entries, which include Mark Twain's *Pudd'nhead Wilson*, George Washington Cable's *Old Creole Days*, Hamlin Garland's *A Spoil of Office* (subtitled "A Story of the American West"), and Booth Tarkington's *A Gentleman from Indiana*. Unlike the text-only case, this regionalist cluster, in addition to being strongly split by gender, does not stretch to include later science fiction, the most probable explanation being that the direct use of publication dates as a feature discourages aggregation across longer time spans. Depending on one's views concerning the proper historical limits of individual genres, this fact may represent a bug or a feature of the extended model. I'm inclined to see it as valuable in most instances, representing more closely existing mainstream critical practice with respect to periodization. But it's also a reminder that feature selection matters; computational models work on the data we engineer for them, and our engineering is part of the interpretive process because it embeds assumptions about the phenomena in question.

So we have, according to the full-data model, female-authored sentimental fiction and male-authored regionalism in the older reaches of the corpus, and gendered varieties of detective and mystery fiction toward the contemporary end of the spectrum. Science fiction is in this case surprisingly absent from the scheme as a genre unto itself, perhaps because its content is insufficiently standardized - and its imagined geography too varied or fanciful - to assort tightly once gender and publication history are introduced as explicit factors. Pieces of sci-fi are, however, rolled together with detective fiction in some instances.[24] In any case, the DBSCAN algorithm applied to this corpus and set of engineered features seems to be good at identifying and anatomizing the varieties of highly typified genre literature that we most associate with the term.

What are we to make, then, of the last large cluster that it produces, which contains more than 170 volumes, exclusively by men, including many of the most canonical living or recently-deceased writers of serious literary fiction? The group in question - cluster 8 - covers books by John Steinbeck, Saul Bellow, John Updike, John Cheever, Richard Yates, Kurt Vonnegut, Don DeLillo, Padgett Powell, Tim O'Brien, Gil Scott-Heron, Paul Theroux, Walker Percy and, for good measure, James Michener, George R. R. Martin, Stephen King, L. Ron Hubbard, and Elie Wiesel. If we hadn't just walked through the rest of the full-data and text-only models, one might suspect a simple mistake; no critic, I think it's safe to say, would identify this group of writers as working in a shared genre, nor, indeed, as having much in common beyond the facts of their demography and renown. But they do indeed share notable features: they write books of above-average length, devote more than average attention to the American West (in part because they write later in the twentieth century, when attention in general shifts westward), and revive aspects of older fiction (higher reading levels, less aversion to description

[23] It may be helpful to think of the clusters identified here as similar to galaxies: rare, high-density agglomerations embedded in mostly empty space.

[24] Compare Underwood's results in "The Life Cycles of Genres" Underwood investigates science fiction using supervised methods and a different feature set, finding important signs of homogeneity within that genre.

and emotion) toward the end of the century.

It's less interesting, however, to characterize a set of books that are already well known than it is to think about the implications of their algorithmic identifiability and about the canonical contemporary fiction that isn't included here. It's no surprise, of course, to find that a version of the postwar canon is dominated by dead white men. Yet the writers identified in this case are usually thought to have produced quite varied texts, notwithstanding a few identifiable subgroups of the Updike-Cheever-Yates sort. It is striking, then, to see them sort together via a method that specifically excludes groups heterogeneous in content and form. A method, in fact, that recognizes the significant diversity of much modernist literature, of contemporary romance fiction, and of most Af-Am lit - finding them insufficiently uniform to constitute any high-density cluster - but is really good at identifying codified genre fiction. If we believe the results, it is hard not to conclude that many "serious" contemporary white male writers constitute a group both as internally consistent and as finally atypical of the larger literary field as are Sue Grafton mysteries or late-nineteenth-century sentimental and didactic fiction.

Still, cluster 8 is a long way from representing the full postwar canon. What's missing from this genre-like group? Women and writers of color (Gil Scott-Heron excepted), obviously; Toni Morrison, Alice Walker, Ralph Ellison, Ishmael Reed, Samuel Delaney, and James Baldwin are all excluded, as are Margaret Atwood, Eudora Welty, Danielle Steele, Jackie Collins, Joyce Carol Oates, Louise Erdrich, Flannery O'Connor, Joan Didion, Mary Gordon, Barbara Kingsolver, and Pearl S. Buck. A few pretty canonical white male authors are absent, too, including Thomas Pynchon, E. L. Doctorow, Mario Puzo, Norman Mailer, and Philip Roth. Text by all of these authors are present in the corpus, but absent from cluster 8 and, indeed, from nearly all of the other clusters.

There are two conclusions one might draw from the observed pattern of inclusion in (and exclusion from) cluster 8. On one hand, it is possible to emphasize the coherence of this group of texts and to see it as supporting an older and more conservative view of literary value. This approach would note that many of the books that critics have treated as worthy of academic attention occupy a distinct generic space, that they are, in short, *different* from the large majority of twentieth-century fiction. "Better," perhaps, if one works in those terms, though a hypothetical conservative critic would need to take L. Ron Hubbard and George R. R. Martin along with Bellow and Updike.

On the other - and, I believe, preferable - hand, one might emphasize the specifically *generic* coherence of the texts in cluster 8, a coherence that argues against the anti-generic interpretation of literary fiction. On this reading, the remarkable, utterly surprising generic affinity of a large group of highly respected and seriously studied (mostly) dead (mostly) white men suggests the need for a reconsideration of variety and diversity as purported hallmarks of contemporary literary fiction. Critics and scholars, this view argues, would no more limit their professional purview to detective fiction by claiming that it is the core of contemporary literary production than they ought to go on treating Steinbeck, Updike, Vonnegut, DeLillo, and O'Brien as if those writers were, indeed, figures typical of our literary moment. They are not; they are representatives of a single, atypical, highly internally homogeneous group of writers and texts.

Yet one can imagine a syllabus under the heading "Contemporary American Fiction" including none but texts from cluster 8; it would be limited and unsatisfactory in obvious ways, but it wouldn't be a joke or a head-scratching error in the way that listing only Sue Grafton and Marcia Muller - or Philip K. Dick, Elmore Leonard, and Robert A. Heinlein - would be. The sooner we realize that these groups are of the same sort as far as generic specificity is concerned, the better will be our understanding of the system of contemporary fiction writ large and the better will we be able to pursue a truly diverse, inclusive, and intellectually ambitious formulation of the literary field.

A final point bears mentioning. The differences between the text-only and full-data models are modest but, in some instances, important. The text-only model suggests a continuity between early regionalism and postwar science fiction that the full-data model does not detect. The full-data model highlights tendencies toward historical and gender specificity within genres that the text-only model contains but does not foreground. And the full-data model helps us see and understand the stubborn exclusion of women and minorities from the contemporary literary canon. Adding a limited amount of social and historical information to the otherwise strictly textual feature set produces output that, unsurprisingly, enlarges the historical and social skews that are detectable using textual features alone. These differences of emphasis are neither good nor bad in isolation, though they do increase our confidence that the method is behaving as expected by shifting its output in response to changing input. We should be clear, though, that the incorporation of extra-textual data does not introduce a distortion of an otherwise true or correct textual model, unless one believes *as an interpretive stance* that texts ought to be studied in isolation from the circumstances of their production and reception - a position that has nothing to do with the use of computation as an aide to literary analysis and one at odds with most contemporary critical

practice.

In the case of genre in twentieth-century American fiction, unsupervised learning methods applied to mixed thematic, formal, and bibliographic textual features have allowed us to confirm the identifiability and coherence of recognizably generic fiction in a large literary corpus. While we do not yet see strong evidence of a *Program Era*-like division of specifically literary fiction into a handful of well-defined subgenres, the methods employed have lead to the surprising discovery of a genre-like cluster of late-century novels by prominent, almost uniformly white, male writers. This last cluster sheds light on the extent to which the contemporary fiction canon functions as a genre unto itself and suggests the continuing need for a more expansive treatment of literary production in the twentieth century and beyond if critics' assessments are to reflect sociotextual artifacts beyond the limits and conventions of a single genre.